Table of Contents

[Exploratory Data Analysis 2](#_Toc193826537)

[Project Description: 2](#_Toc193826538)

[Project Title: 3](#_Toc193826539)

[Objective: 3](#_Toc193826540)

[Dataset: 3](#_Toc193826541)

[Methodology: 4](#_Toc193826542)

[Tools and Technologies: 9](#_Toc193826543)

[Descriptive Analysis 9](#_Toc193826544)

[Project Description: 9](#_Toc193826545)

[Project Title: 9](#_Toc193826546)

[Objective: 9](#_Toc193826547)

[Dataset: 9](#_Toc193826548)

[Methodology: 10](#_Toc193826549)

[Tools and Technologies: 18](#_Toc193826550)

[Deliverables: 18](#_Toc193826551)

[Data Wrangling 19](#_Toc193826552)

[Project Description: 19](#_Toc193826553)

[Project Title: 19](#_Toc193826554)

[Objective: 19](#_Toc193826555)

[Background: 19](#_Toc193826556)

[Dataset: 19](#_Toc193826557)

[Methodology: 20](#_Toc193826558)

[Tools and Technologies: 22](#_Toc193826559)

[Deliverables: 22](#_Toc193826560)

[Timeline: 22](#_Toc193826561)

[Data Quality Control 22](#_Toc193826562)

[Project Description: 22](#_Toc193826563)

[Project Title: 22](#_Toc193826564)

[Objective: 22](#_Toc193826565)

[Scope: 22](#_Toc193826566)

[Methodology: 28](#_Toc193826567)

[Deliverables: 29](#_Toc193826568)

[Timeline: 29](#_Toc193826569)

# Exploratory Data Analysis

Project Description: Exploratory Data Analysis (EDA) on DAP processing and summarizing bulk datasets required by the City of Vancouver

Project Title: AWS Data Analytic Platform for The City of Vancouver

Objective: The primary goal of this project is to perform exploratory data analysis (EDA) on the 311 contact center metrics dataset to design and implement a data analytic platform to support descriptive analysis. By analyzing various features such as Date, Calls\_Offered, Calls\_Handled, Calls\_Abandoned, Average\_Speed\_of\_Answer, Service\_Level, and BI\_ID, we aim to understand the factors that influenced the likelihood of the service of 311.

Dataset: This dataset consists of 311 service information, including details such as:

* Date: 2024-1-1 to 2025-1-31
* Calls\_Offered: the time of 311 offering calls (seconds)
* Calls\_Handled: the time of 311 handling calls (seconds)
* Calls\_Abandoned: the time of 311 abandoning calls (seconds)
* Average\_Speed\_of\_Answer: the time of average speed of answers from 311 (seconds)
* Service\_Level: 311 service level (0-1)
* BI\_ID: The ID number of providing services

When uploading the dataset in AWS, the data profile overview (see Figure 1) can show the summary of this dataset.

***Figure 1: Data Profile Overview***

A screenshot of a computer

AI-generated content may be incorrect.

## Methodology:

1. Data Collection and Preparation and Cleaning:

* Download the dataset from the City of Vancouver
* Observing the initial dataset, which includes missing values, outliers, and data type.
* The initial data is mostly cleaning, so I just adjusted the two decimal places and then changed this dataset to a CSV file to be uploaded to AWS S3 Buckets (see figure 2 and figure 3)

***Figure 2: S3 Bucket Building***

A screenshot of a computer

AI-generated content may be incorrect.

***Figure 3: Data Cleaning***

A screenshot of a computer

AI-generated content may be incorrect.

2. Descriptive Statistics:

All data are numerical values, but Service\_Level and Average\_Speed\_of\_Answer are not integers.

3. Data Cataloging:

I created a new bucket called 311-ctc-mtr-cur and sub-folders for metrics including system and user folders, which will be stored for summarization. When the Crawler was ready, the cleaning data was transferred to databases for 34 cataloging and prepared for summarization (See Figure 4, Figure 5, Figure 6).

***Figure 4: Metrics Folders Building***

A screenshot of a computer

AI-generated content may be incorrect.

***Figure 5: Crawlers Building***

A screenshot of a computer

AI-generated content may be incorrect.

***Figure 6: Transfer Cleaning Data to Databases for Catalog***

A screenshot of a computer

AI-generated content may be incorrect.

4. Data Summarization:

Two ETL pipelines were established, and the metrics report was for systems and users.

***Figure 7: ETL Pipeline***

A screenshot of a computer

AI-generated content may be incorrect.

***Figure 8: Metrics Report***

A screenshot of a computer

AI-generated content may be incorrect.

5. Conclusion:

This project designed and implemented a data analytic platform that supports descriptive analysis and provided a thorough report that breaks down each step of the process. City of Vancouver can leverage AWS to drive data-based decision-making with minimal cost and high efficiency

6. Data Cleaning (in-class activity):

***Table 1: Poor Data Quality Dataset***

A screenshot of a computer screen

AI-generated content may be incorrect.

In this class activity, we practiced identifying the poor data quality including duplicate, missing values, invalid values, inconsistent format, unbalanced, typo, encoding, and so forth.

## Tools and Technologies:

AWS S3, Glue, Glue DataBrew

# Descriptive Analysis

Project Description: Descriptive Analysis of Sales Performance (W6 in-class)

Project Title: Understanding the Sales Performance

Objective: Questions about Sales Performance Analysis, Impact of Marketing Spend, Effect of Temperature on Sales, and Website Traffic Insight were answered.

Dataset: The dataset includes the following key features:

* Date: 2023-1-1 to 2024-3-31
* Sales: Sales of each corresponding date
* Temperature: the temperature of each corresponding date
* Marketing Spend: the expenditure on the marketing of each corresponding date
* Customer purchase behavior: there are three categories, including A,B, and C.
* Website Traffic: the amount of traffic of each corresponding date
* Outlier Flag: there are two

***Table 2: Sales Performance Dataset***

A screenshot of a computer

AI-generated content may be incorrect.

## Methodology:

1. Data Collection and Preparation:

* Load the dataset using data analysis tools, including Excel and SQL
* The original data is cleaned.

2.1 Descriptive Statistics:

***Table 3: Questions about Sales Performance Analysis, SQL Syntax, and Analysis***

A screenshot of a computer

AI-generated content may be incorrect.

Calculate summary statistics for key variables, including:

* Average sales
* The highest and lowest sales
* How do sales vary by date

3.1 Data Visualization:

***Table 4: Data Visualization by Excel***

A table with numbers and numbers

AI-generated content may be incorrect.

Create visual representations to illustrate findings:

* Average sales: 139.0625
* The highest sales: 300
* The lowest sales: 50
* Sales vary in each date

2.2 Descriptive Statistics:

***Table 5: Questions about Impact of Marketing Spend, SQL Syntax, and Analysis***

A screenshot of a computer

AI-generated content may be incorrect.

Calculate summary statistics for key variables, including:

* Correlation between marketing spend and sales
* The correlation between marketing spend and website traffic
* Which dates had the highest and lowest marking spending

3.2 Data Visualization:

***Table 6: Data Visualization by Excel***

A screenshot of a graph

AI-generated content may be incorrect.

Create visual representations to illustrate findings:

* Correlation between marketing spend and sales: 0.313434298
* Correlation between marketing spend and website: 0.318149461
* The highest marketing spend date: 2023-3-15
* The lowest marketing spend date: 2023-1-1

2.3 Descriptive Statistics:

***Table 7: Questions about Effect of Temperature on Sales, SQL Syntax, and Analysis***

A screenshot of a computer

AI-generated content may be incorrect.

Calculate summary statistics for key variables, including:

* Correlation between temperature and sales
* Does higher temperatures lead to more or fewer sales?

3.3 Data Visualization:

***Table 8: Data Visualization by Excel***

A screenshot of a graph

AI-generated content may be incorrect.

Create visual representations to illustrate findings:

* Correlation between temperature and sales: 0.701793526
* When the temperature is higher, the average sales are higher.

2.4 Descriptive Statistics:

***Table 9: Questions about website traffic insights, SQL Syntax, and Analysis***

A screenshot of a computer

AI-generated content may be incorrect.

Calculate summary statistics for key variables, including:

* What is the average daily website traffic?
* On which dates was the website traffic highest and lowest?
* Does an increase in website traffic lead to higher sales?

3.4 Data Visualization:

***Table 10: Data Visualization by Excel***

A screenshot of a graph

AI-generated content may be incorrect.

Create visual representations to illustrate findings:

* The average daily website traffic: 985.625
* The highest website traffic is 2500 on 2023-2-29
* The correlation between website traffic and sales: 0.224236537

4. Insights and Findings:

* The highest correlation of sales is the temperature, and the higher the temperature is, the higher the sales are.
* The highest sales happened in March, so marketing spending was also used most in this month.

5. Recommendations:

* The sales rely more on the temperature, so adjusting the suitable temperature in the store is essential, although the outside temperature cannot be controlled.
* The marketing expenditure can be reduced gradually, particularly in the month with higher sales.

6. Other Descriptive Analysis (Weekly Activity 6)

***Figure 9. Using Athena to Do descriptive Analysis***

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

## Tools and Technologies:

* SQL and Excel for data analysis, and AWS Athena for descriptive data analysis
* Data visualization tool is Excel

## Deliverables:

* A detailed report summarizing the methods, findings, and recommendations.
* Visualizations and dashboards to present key insights clearly.
* A presentation for stakeholders to communicate important findings and suggestions for future action.

This descriptive analysis project aims to comprehensively understand sales performance, the impact of marketing expenditures, the impact of the temperature on sales, and the website traffic insights.

# Data Wrangling

Project Description: The City of Vancouver needs to migrate to AWS, and we implemented a data analytic platform.

Project Title: AWS Data Analytic Platform for The City of Vancouver

Objective: The primary goal of this project is to perform comprehensive data wrangling to prepare a robust dataset for the 311 Contact Center at The City of Vancouver. By cleaning, transforming, and consolidating data from various sources at AWS, the project aims to enhance the accuracy and usability of this data for subsequent analysis and reporting.

Background: The City of Vancouver needs to migrate its data to AWS, and I selected one of the datasets about the 311 Contact Center. However, this data needs cleaning, otherwise making it challenging to derive meaningful insights. Effective data wrangling will facilitate better decision-making.

Dataset: The data wrangling process will involve one dataset, including:

* Date: 2024-1-1 to 2025-1-31
* Calls\_Offered: the time of 311 offering calls (seconds)
* Calls\_Handled: the time of 311 handling calls (seconds)
* Calls\_Abandoned: the time of 311 abandoning calls (seconds)
* Average\_Speed\_of\_Answer: the time of average speed of answers from 311 (seconds)
* Service\_Level: 311 service level (0-1)
* BI\_ID: The ID number of providing services

## Methodology:

1. Data Collection:

* Download the dataset from the City of Vancouver
* Observing the initial dataset, which includes missing values, outliers, and data type.

2. Data Assessment and Cleaning:

***Figure 10. Quality Check in Visual ETL***

A screenshot of a computer

AI-generated content may be incorrect.

I did the ruleset for data quality and then set the conditional router, including passed and failed, and then changed the schema before load, including dropping the useless columns. Finally, I loaded the Passed and Failed into the corresponding folders.

3. Data Transformation and Consolidation:

***Figure 11: Failed Dataset in Transfer Bucket***

A computer screen with a white background

AI-generated content may be incorrect.

***Figure 12: Passed Dataset in Transfer Bucket***

A screenshot of a computer

AI-generated content may be incorrect.

After loading the passed and failed in Visual ETL, the corresponding datasets have already been published into each sub-bucket in the transfer bucket, which can be downloaded for different uses.

## Tools and Technologies:

* AWS S3 and AWS Glue

## Deliverables:

* A cleaned and transformed 311 Contact Center dataset ready for analysis, particularly for Passed Dataset in a CSV format.
* Confirmation of data quality checks conducted during the process.

## Timeline:

* Expected completion of the project: Week 9 and Week 10, including phases for assessment, cleaning, transformation, and documentation.

This data wrangling project aims to establish a high-quality dataset that enables the City of Vancouver to conduct adequate 311 Contact Center Data.

# Data Quality Control

Project Description: The HR Department of UCW implemented the Data Quality Control

Project Title: Implementation of Data Quality Control Measures HR Department of UCW

Objective: The primary objective of this project is to establish a comprehensive Data Quality Control (DQC) framework at AWS for the HR department of UCW. This framework will ensure the organization's data's accuracy, completeness, consistency, and reliability, enhancing decision-making processes and overall business performance.

Scope: The project will focus on the following key areas:

* Data Profiling: Analyzing existing datasets to assess quality levels.

***Figure 13: Data Profiling for HR of UCW***

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

* Data Cleansing: Developing processes to correct inaccuracies and eliminate duplicates.

***Figure 14: Data Cleaning for HR of UCW***

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

* Data Validation: Implementing validation rules and checks to ensure data integrity.

***Figure 15: Data Validation for HR of UCW***

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

* Monitoring and Reporting: Establishing ongoing monitoring processes and dashboards to track data quality metrics.

***Figure 16: Data Monitoring for HR of UCW***

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

## Methodology:

1. Data Profiling:

* Utilize data profiling tools to assess the quality of identified datasets, focusing on completeness, uniqueness, validity, consistency, and accuracy.
* Document findings to highlight areas requiring immediate attention.

2. Data Cleansing Processes:

Develop and implement procedures for data cleansing, which includes renaming.

3. Validation Rules and Procedures:

* Set up validation rules for new data entries to reduce the risk of poor-quality data being introduced into the system.
* Create data entry guidelines to promote consistency and accuracy.

4. Monitoring:

* Implement monitoring tools and dashboards that provide real-time data quality metrics and alerts for significant deviations.

Tools and Technologies:

* AWS CloudWatch
* WAS Glue DataBrew

Deliverables:

* A comprehensive Data Quality Control plan detailing processes, metrics, and responsibilities.
* Cleaned and validated datasets ready for analysis and reporting.
* A monitoring dashboard that visualizes data quality metrics in real time.

## Timeline:

Expected completion of the project: Week 3, Week 4, and Week 9.

This Data Quality Control initiative aims to empower UCW to enhance its HR data integrity and reliability, resulting in improved decision-making, operational efficiency, and compliance with regulatory requirements.